Two-Component Multicanonical Monte Carlo Method for Effective Conformation Sampling

JUNICHI HIGO,[1] NOBUYUKI NAKAJIMA,[1] HIROKI SHIRAI,[1] AKINORI KIDERA,[2] HARUKI NAKAMURA[1]

[1] *Department of Bioinformatics, Biomolecular Engineering Research Institute (BERI), 6-2-3, Furuedai, Suita, Osaka 565, Japan*
[2] *Faculty of Science, Kyoto University, Kyoto, Japan*

Received 19 February 1997; accepted 13 August 1997

ABSTRACT: A multicanonical algorithm, which is one of the most powerful conformation-sampling methods to obtain the density of states described by a component (i.e., the total potential energy), was extended to obtain the density of states described by two components. This method was tested on a simplified model for bacteriorhodopsin, which is a membrane protein consisting of seven helices. Two kinds of simulation were done by adopting different sets of two components: in one set, the components were the site-specific and the non–site-specific energies between helices; and, in the other set, the total potential energy and the end-to-end distance (distance between the first and the seventh helices) were used. In both simulations, a wide and flat probability distribution was obtained, showing the efficiency of the two-component method. A variety of applications may be possible by effectively selecting the two components, such as the van der Waals and electrostatic energies, the intermolecular and intramolecular interactions, the solute–solute and solute–solvent interactions, or an energy and a reaction coordinate.
J Comput Chem **18**: 2086–2092, 1997

Keywords: conformation sampling; density of states; multicanonical; Monte Carlo simulation; probability distribution

Correspondence to: J. Higo

CCC 0192-8651 / 97 / 162086-07

Introduction

Effective conformation sampling (or configuration sampling) is important in the study of systems that involve a large number of local energy minima on the potential energy surface. In conventional simulation techniques, conformation trapping frequently occurs and the sampling is restricted within a small region on the potential surface. Recently, some effective sampling methods have been developed. The multicanonical algorithm[1–3] provides a wide energy distribution and yields a generalized ensemble, from which a density of states and an ensemble at any temperature can be derived. This algorithm was used for Monte Carlo[4–10] and molecular dynamics[11,12] simulations, and has been applied to a spin system,[4,5] a hard-sphere system,[6,7] and a biological system.[8–12] Simulated tempering[13] and an 1/k algorithm[14] can also be introduced for effective sampling. It has been reported that these three methods are closely related and that the difference in the sampling efficiencies among the three is only marginal.[15,16]

In some systems, dividing the total energy into the energy elements (e.g., van der Waals + electrostatic, or intramolecular + intermolecular) makes the problem easier and more clear. Accordingly, it is useful to express the density of states by these elements. Together with an energetic element, the introduction of a structural parameter, such as a reaction coordinate, is also useful for describing the density of states. Kumar et al. performed simulations of polypeptides, yielding a free-energy surface (or potential of mean force) that was multidimensionally described.[17] Their method was an extension of the weighted histogram analysis method.[18,19]

We extended the multicanonical method to obtain the density of states that is described by two elements (i.e., two components), and tested the method on bacteriorhodopsin, a membrane protein consisting of seven helices. Suwa et al.[20] simplified this molecule by replacing each helix with a cylinder, where the change of the helix orientation was discrete. In our study, each helix was represented by a disk, and the system was reduced to a two-dimensional model. The change of the helix orientation was continuous. Two kinds of simulation were done by adopting different sets of two components: In one, the two components were the site-specific and the non–site-specific interactions between helices (the sum of the two was the total potential energy). Therefore, both components were energetic. In the other simulation, the two components were the total potential energy and the end-to-end distance (the distance between the first and the seventh helices). Here, one component was energetic and the other was structural. We discuss the efficiency and the significance of this new method.

Methods

MULTICANONICAL ALGORITHM

The thermodynamic properties of a system at equilibrium are determined by the density of states, $n(E)$, where the number of states (i.e., conformations or configurations) within a potential-energy window between E and $E + dE$ is given by $n(E)\,dE$. In the multicanonical procedure, $n(E)$ has been obtained through an iteration of simulations.[1–3] Here, we denote this method as a one-component method, because the density of states obtained is described by one component (i.e., E). We now briefly explain the one-component method with a Monte Carlo scheme.

First, a pre-run is done with the conventional Monte Carlo method.[21] The density of states, $n_0(E)$, is then defined as

$$n_0(E) = P_0(E)\exp(E/T) \tag{1}$$

where $P_0(E)$ is the probability distribution (as a function of energy) obtained from the pre-run, and T is the temperature (here, the Boltzmann or gas constant is set to unity). In the pre-run, the temperature is set at a high value to insure that the conformation can overcome potential-energy barriers, and then $n_0(E)$ is evaluated only in a high energy region.

Next, the first multicanonical simulation starts, using the function $n_0(E_{\mathrm{acp}})/n_0(E_{\mathrm{try}})$ for the judgment of acceptance (E_{acp} is the energy of the conformation accepted at a current step of the simulation and E_{try} is that of a trial conformation): If $n_0(E_{\mathrm{acp}})/n_0(E_{\mathrm{try}}) \geq 1$, accept the trial. If $1 > n_0(E_{\mathrm{acp}})/n_0(E_{\mathrm{try}}) \geq \mathrm{ran}[0,1]$, accept the trial. Otherwise, reject. Here, ran$[0,1]$ is a random number in the range $[0,1]$. Note that using $n_0(E_{\mathrm{acp}})/n_0(E_{\mathrm{try}})$ in the judgment corresponds to using $n_0(E)^{-1}$ instead of $\exp[-E/T]$ for the statistical weight during the simulation. By this procedure, a

flat probability distribution [i.e., $P_1(E) = \text{const.}$] is obtained if $n_0(E)$ has been correctly estimated. However, in general, $P_1(E)$ is not completely flat, because $n_0(E)$ is approximated from a pre-run of a finite length at a high temperature. From this simulation, the density of states $n_1(E)$ is approximated as:

$$n_1(E) = P_1(E)n_0(E) \tag{2}$$

Then, the second multicanonical simulation starts, using the function $n_1(E_{\text{acp}})/n_1(E_{\text{try}})$ for the judgment of acceptance, and yields $P_2(E)$ and $n_2(E)$ where the relation $n_2(E) = P_2(E)n_1(E)$ is used. Through a series of multicanonical simulations, the density of states becomes more accurate. Remember that $n_0(E)$ was obtained only in a high-energy region. In the series of multicanonical simulations, the density of states is gradually extrapolated to a low-energy region.

Now, we introduce a two-component, multicanonical method. Let us denote the density of states described by two components as $n(\alpha, \beta)$, where α and β are arbitrary. When $\alpha = \beta = E$, this density of states becomes $n(E)$. In this study, α was an energetic component, and β was either energetic or structural. In the pre-run, the probability distribution described by the two components, $P_0(\alpha, \beta)$, was calculated by counting the conformations visiting a two-component window between α and $\alpha + d\alpha$ and between β and $\beta + d\beta$. Then, the density of states, $n_0(\alpha, \beta)$, was written in the same manner in eq. (1), as.

$$n_0(\alpha, \beta) = P_0(\alpha, \beta) \exp(E/T) \tag{3}$$

The first multicanonical simulation followed, using the function $n_0(\alpha_{\text{acp}}, \beta_{\text{acp}})/n_0(\alpha_{\text{try}}, \beta_{\text{try}})$ for the judgment of acceptance, where α_{acp} and β_{acp} are the two components for the currently accepted conformation, and α_{acp} and β_{acp} were those for the trial. The density of states at this stage was estimated by:

$$n_1(\alpha, \beta) = P_1(\alpha, \beta)n_0(\alpha, \beta) \tag{4}$$

In a series of successive simulations, the two-component density of states from the ith multicanonical simulation was calculated iteratively as:

$$n_i(\alpha, \beta) = P_i(\alpha, \beta)n_{i-1}(\alpha, \beta) \tag{5}$$

Finally, the probability distribution function became flat in a wide two-component range. In the current study, the temperature, T, was set to 1.0 in the pre-run.

A SIMPLIFIED SYSTEM

We made a simplified model of bacteriorhodopsin, which is a membrane protein composed of seven helices connected by loops.[22–24] The helical axes are approximately perpendicular to the membrane. In this study, a two-dimensional model was prepared, in which each helix was represented by a rigid disk (Fig. 1). Several interaction sites were placed on the periphery of the disk. Two types of attractive interactions, the site-specific (E_{ss}) and the non−site-specific (E_{ns}) ones, were introduced:

$$E_{ss} = f_{ss}\Sigma(r_{ij}/r_0 - 1) \tag{6a}$$

$$E_{ns} = f_{ns}\Sigma(R_{kl} - R_0)/(R_{kl} - R_{v\,dw}) \tag{6b}$$

Here, r_{ij} is the distance between two interaction sites (denoted by i and j) on different disks, r_0(= 3) is the radius of effective site-specific interaction ($E_{ss} = 0$ for $r_{ij} > r_0$), R_{kl} is the distance between two disks denoted by k and l, R_0 (= 18) is the radius of non-site-specific interaction ($E_{ns} = 0$ for $R_{kl} > R_0$), and $R_{v\,dw}$ is the sum of the radii of two disks (= 12 = 6 + 6, where the value of 6 corresponds to the radius of a helical wheel). Figure 1 shows these parameters. The contribution coefficients, f_{ss} and f_{ns}, are set to −1.7 and −0.42, respectively. All of the parameters and physical quantities such as energy, T, and length, were treated as dimensionless variables. The summation in eq. (6a) was taken over all pairs of interaction sites on different disks, and the summation in eq. (6b) was over all pairs of disks. We determined the positions of the interaction sites so that the experimentally determined packing of the helices[22–24] is in the global energy minimum, as shown in Figure 1b. We supposed that the length of the loops connecting the helices was 14. The effect of these loops on the potential energy was counted by rejecting trial conformations only when $R_{kl} > 14$ in the simulation. The trial was also rejected when the disks overlapped one another (i.e., $R_{kl} < R_{v\,dw}$). Thus, the total potential energy was defined as $E = E_{ss} + E_{ns}$.

The number of local minima was roughly estimated to be at least 8×10^4. Here, we assumed that the helices bound one another in linear conformations: Because the first helix has two interaction sites, the number of ways for this helix to bind to the second helix is two. Similarly, the number of ways for the last helix to bind is four. The second helix has three sites, and so the number of ways to bind to the first and third helices is six (= 3 × 2).

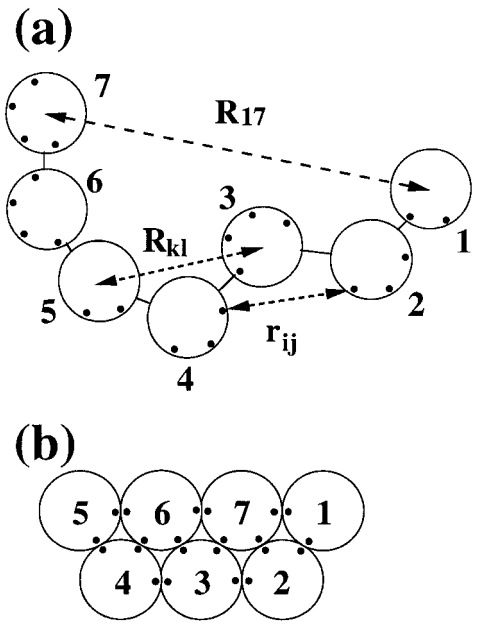

FIGURE 1. Model of bacteriorhodopsin. Helices, represented by large open circles, are numbered from the N-terminus. Small, filled circles are interaction sites. In the random conformation (a), broken lines represent r_{ij} and R_{kl} used in eq. (6), as well as the end-to-end distance R_{17}. Solid lines between helices represent loops. The global energy–minimum conformation (b) corresponds to an experimentally obtained conformation.

The numbers of ways for the other helices, except for the first and last helices, were estimated similarly. Finally, a value 82,944 ($= 2 \times 6 \times 12 \times 6 \times 2 \times 12 \times 4$) was obtained. In this estimation, branched or ring-closured conformations were not counted, and the excluded volume effect was not considered.

The molecule was so highly simplified that the results from this study have no biophysical significance. The aim of this study is to propose a new conformation-sampling method

TWO COMPONENTS

Two kinds of multicanonical simulations were done by adopting different sets of components: One set was E_{ss} and E_{ns}, and the other was E and R_{17} ($= |\mathbf{R}_{17} - \mathbf{R}_7|$; the end-to-end distance is shown in Fig. 1a). We called the former set ''energy–energy'' and the latter ''energy–distance.''

Because the density of states was defined in the two-component space, the sampling can be done in a selected region in the space. For instance, to obtain the global energy–minimum conformation, there are several pathways to decrease the energy. In a simulation, when a trial conformation was generated outside of the selected region, the trial was rejected. In this study, regions with low E_{ss} and high E_{ns} in the energy–energy set and regions with low E and long R_{17} in the energy–distance set were eliminated in the sampling procedure, as shown schematically in Figure 2.

In the simulation of the energy–distance set, we calculated the free energy dependence on R_{17}:

$$G(R_{17}, T) = -T \ln[Z(R_{17}, T)] \quad (7)$$

where:

$$Z(R_{17}, T) = \int n(E, R_{17}) \exp(-E/T)\, dE \quad (8)$$

Note that the entire free energy is not given solely by the summation of $G(R_{17}, T)$.

Results

The $P(E_{ss}, E_{ns})$ obtained from the last simulation of the energy–energy set is shown in Figure 3, and reveals a flat distribution. The region with log $P = 0$ is the eliminated region from the sampling. Note that the sampling was not done in the regions with low E_{ss} and high E_{ns}. The bin size was 0.25 along both the E_{ss} and E_{ns} axes. The number of total bins was 139. The number of iterations of the multicanonical simulations was 15, with 2×10^7 steps for each, except for the last simulation of 1×10^8 steps. The lowest energy conformation during the simulation corresponded to the conformation with the global minimum energy. Some of

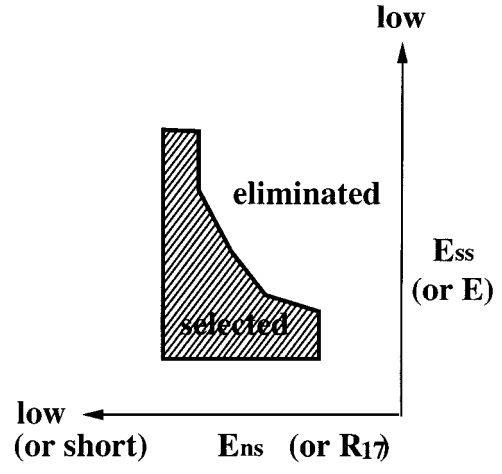

FIGURE 2. Scheme for the selected and rejected regions in the sampling.

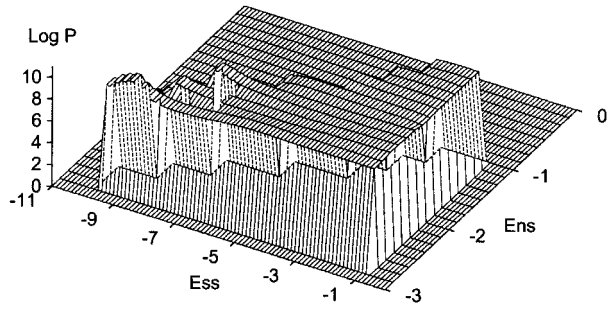

FIGURE 3. Probability distribution $P(E_{ss}, E_{ns})$. The distribution is given as log-scale and normalized so that the maximum P was set to 10. The region with log $P = 0$ was eliminated from the sampling. Broken lines represent hidden sides.

the sampled low-energy conformations are shown in Figure 4. The smallest density of states (i.e., $n_{\min}$), in which the global energy–minimum conformation was involved, was about 10^{+22} times smaller than the largest (i.e., $n_{\max}$): $n_{\max}/n_{\min} = 10^{+22}$.

The $P(E, R_{17})$ obtained from the last simulation of the energy–distance set is shown in Figure 5, and also displays a flat distribution. The region with log $P = 0$ is the eliminated region from the sampling. Note that the sampling was not done in the regions with low E and long R_{17}. The bin size was 0.25 along the E axis and 0.84 along the R_{17} axis. The number of total bins was 269. The number of iterations was ten, taking 2×10^7 steps for each, except for the last simulation of 1×10^8 steps. Profiles of E and R_{17} are presented in Figure 6, with some snapshots of the conformations. In the first part of Figure 6, a correlation between

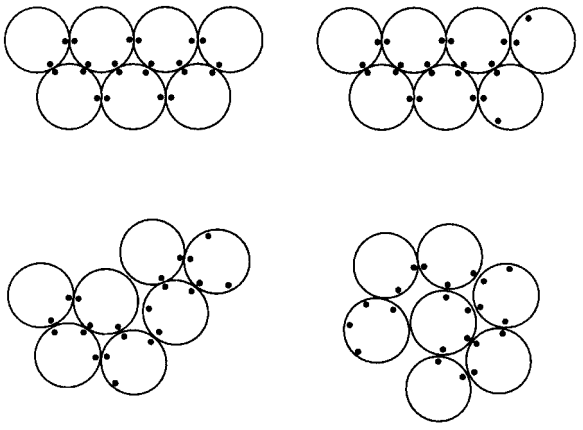

FIGURE 4. Sampled low-energy conformations, which contain the global energy-minimun conformation.

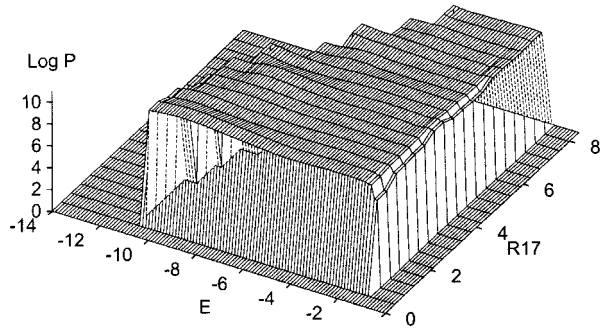

FIGURE 5. Probability distribution $P(E, R_{17})$. See the legend to Figure 3.

E and R_{17} can be seen, which will be discussed. The free energy dependence on R_{17} is shown in Figure 7. This figure shows the reasonable property that random conformations are stable at high temperature, whereas packed states are stable at low temperature.

Discussion

The two-component, multicanonical method includes more information than the one-component method. It is possible to obtain the free energy dependence on each component, as shown in Figure 7. If a reaction coordinate is relevantly selected depending on a system, a free energy barrier may be estimated along the coordinate. The evaluation of the free energy dependence is smoothly con-

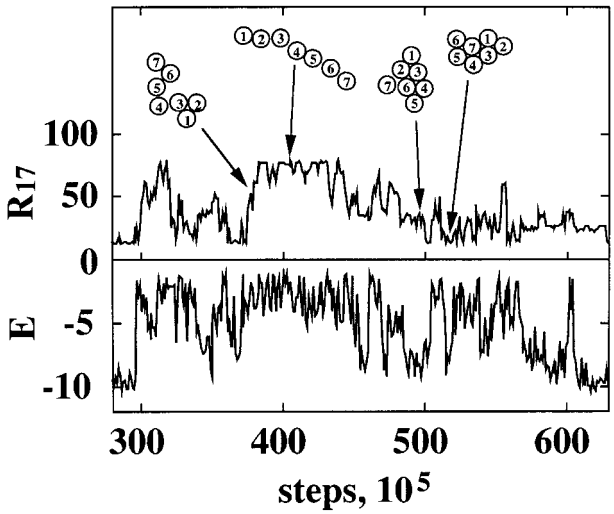

FIGURE 6. Profiles of E and R_{17} with four snapshots, where the digits represent the order of helices from the N-terminus.

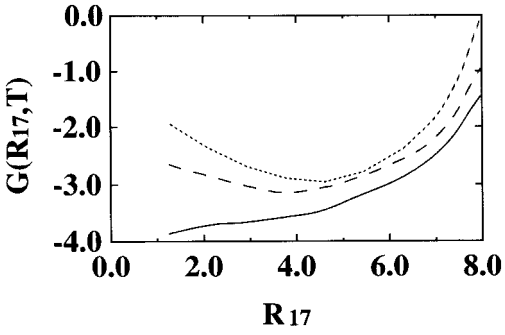

FIGURE 7. Free-energy dependence on R_{17} at three temperatures. Temperatures were 0.05 (solid line), 0.33 (broken line) and 0.5 (dotted line).

nected to the two-component method, as shown by eqs. (7) and (8). Of course, if all of the sampled conformations were stored in a one-component simulation, it would also be possible to derive the dependence; however, it is not practical to do so.

There is a variety of applications for the two-component method that can be performed by selecting a set of two components, such as the van der Waals and electrostatic energies, the inter- and intramolecular interactions, the solute–solute and solute–solvent interactions, or an energy and a reaction coordinate, depending on the individual system.

As a disadvantage of the current method, one may consider that the simulation length of the two-component method should be longer than that of the one-component method, because the sampling must be done in a two-component space. This is true when one wants to evaluate the density of states in a wide region of the two-component space, because it will take a large number of multicanonical iterations to obtain a flat distribution, and thus the one-component method may be more efficient. However, in the two-component method, we can selectively sample the conformation and eliminate the regions that are not important to explore the system, which results in a shorter simulation. The selected region analyzed in this study is schematically drawn in Figure 2. The region with low E_{ss} and high E_{ns} in the energy–energy set and the region with low E and long R_{17} in the energy–distance set have negligibly small densities of states, and so these regions do not contribute to the thermodynamic properties. Therefore, we eliminated these regions in the sampling. The effect of the elimination is shown in the correlation between E and R_{17} in the first part of Figure 6.

Because the aim of this work is to propose a new sampling method, the bin size was not optimized. The flat distribution shown in Figures 3 and 4 will be obtained with fewer iterations when the bin size is more carefully set. It is of interest to determine the optimal bin size for more detailed models.

Selective sampling has another benefit. Suppose that one wants to sample a part of a system (sampling part), while maintaining the conformation of the other part (framework) in an experimentally determined conformation. The introduction of an artificial constraint energy may be a convenient method to achieve this purpose. However, the interpretation of the resultant density of states would be difficult, because the total energy automatically includes the contribution of the additional constraint. In the two-component method, in contrast, the conformation can be constrained without introducing the constraint energy, when multicanonical enhancing is done only for the energy of the sampling part.

Acknowledgments

We thank Drs. J. Sasaki and Y. Kimura from BERI for helpful discussions.

References

1. B. A. Berg and T. Neuhaus, *Phys. Lett.*, **B267**, 249 (1991).
2. B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.*, **68**, 9 (1992).
3. B. A. Berg and T. Celik, *Phys. Rev. Lett.*, **69**, 2292 (1992).
4. K. Rummukainen, *Nucl. Phys.*, **B390**, 621 (1993).
5. J. Lee, *Phys. Rev. Lett.*, **71**, 211 (1993).
6. G. R. Smith and A. D. Bruce, *Phys. Rev. E*, **53**, 6530 (1996).
7. G. R. Smith and A. D. Bruce, *Europhys. Lett.*, **34**, 91 (1996).
8. U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.*, **14**, 1333 (1993).
9. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, **98**, 4940 (1994).
10. A. Kidera, *Proc. Natl. Acad. Sci. USA*, **92**, 9886 (1995).
11. U. H. E. Hansmann, Y. Okamoto, and F. Eisenmenger, *Chem. Phys. Lett.*, **259**, 321 (1996).
12. N. Nakajima, H. Nakamura, and A. Kidera. *J. Phys. Chem.*, **101**, 817 (1997).
13. E. Marinari and G. Parisi, *Europhys. Lett.*, **19**, 451 (1992).

14. B. Hesselbo and R. B. Stinchcombe, *Phys. Rev. Lett.*, **74**, 2151 (1995).
15. U. H. E. Hansmann and Y. Okamoto, *Phys. Rev. E*, **54**, 5863 (1996).
16. U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.*, **18**, 920 (1997).
17. S. Kumar, P. W. Payne, and M. Vásquez, *J. Comput. Chem.*, **17**, 1269 (1997).
18. Z. W. Salsburg, J. D. Jacobson, W. Fickett, and W. W. Wood, *J. Chem. Phys.*, **30**, 65 (1959).
19. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.*, **13**, 1011 (1992).
20. M. Suwa, T. Hirokawa, and S. Mitaku, *Proteins*, **22**, 363 (1995).
21. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.*, **21**, 1087 (1953).
22. R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Dowing, *J. Mol. Biol.*, **213**, 899 (1990).
23. Y. Kimura, D. G. Vassylyev, A. Miyazawa, A. Kidera, M. Matsushima, K. Mitsuoka, K. Murata, T. Hirai, and Y. Fujiyoshi, *Nature* (in press).
24. Y. Kimura, D. G. Vassylyev, A. Miyazawa, A. Kidera, M. Matsushima, K. Mitsuoka, K. Murata T. Hirai, and Y. Fujiyoshi, *Photochem. Photobiol.* (in press).